CLAIMS

1. A method for monitoring a plurality of servers in a cluster and taking corrective action for said servers, said method comprising the steps of:

sending a request to one of said servers, and determining if said one server successfully handles said request and how long it took for said one server to handle said request, and

if a response is received indicating that said one server successfully handled said request, but it took said one server longer than a predetermined time period to handle said request, notifying a dispatcher for said one server to reduce, but not eliminate, a workload of said one server.

2. A method as set forth in claim 1 further comprising the step of:

if said one server does not successfully handle said request within another predetermined time period longer than the first said predetermined time period or does not respond that it handled said request within said other predetermined time period, notifying said dispatcher to remove said one server from said cluster or not to send any subsequent requests to said server.

3. A method as set forth in claim 2 wherein if said one server does not successfully handle said request within said other predetermined time period or does not respond that it handled said request within said other predetermined time period, further comprising the step of automatically sending to said one server a request to restart said one server.

4. A method as set forth in claim 1 wherein said predetermined time period is such that if said one server successfully handles said request, but after said predetermined time period, this indicates that said one server is encumbered or overloaded with requests.

5. A method as set forth in claim 1 wherein the step of sending said request to said one server is performed by sending said request to said one server, bypassing said dispatcher.

6. A system for monitoring a plurality of servers in a cluster and taking corrective action for said servers, said system comprising:

means for sending a request to one of said servers, and determining if said one server successfully handles said request and how long it took for said one server to handle said request, and

means for determining if a response is received indicating that said one server successfully handled said request, but took said one server longer than a predetermined time period to handle said request, and if so, notifying a dispatcher for said one server to reduce, but not eliminate, a workload of said one server.

7. A system as set forth in claim 6 further comprising:

means for if said one server does not successfully handle said request within another predetermined time period longer than the first said predetermined time period or does not respond that it handled said request within said other predetermined time period, and if so, notifying said dispatcher to remove said one server from said cluster or not to send any subsequent requests to said server.

8. A computer program product for monitoring a plurality of servers in a cluster and taking corrective action for said servers, said computer program product comprising:

a computer readable medium;

first program instructions to send a request to one of said servers, and determine if said one server successfully handles said request and how long it took for said one server to handle said request, and

second program instructions to determine if a response is received indicating that said one server successfully handled said request, but took said one server longer than a predetermined time period to handle said request, and if so, notify a dispatcher for said one server to reduce, but not eliminate, a workload of said one server; and wherein

said first and second program instructions are recorded on said medium.

9. A computer program product as set forth in claim 8 further comprising:

third program instructions to determine if said one server does not successfully handle said request within another predetermined time period longer than the first said predetermined time period or does not respond that it handled said request within said other predetermined time period, and if so, notify said dispatcher to remove said one server from said cluster or not to send any subsequent requests to said server.

10. A method for monitoring a plurality of servers in a cluster and taking corrective action for said servers, said method comprising the steps of:

specifying a number of consecutive requests that can be sent to a server and not handled by said server within a specified time period for each of said requests, said number indicating that said server is down;

sending a request to one of said servers, determining that said one server did not successfully handle said request within said specified time period, determining that said number has not yet been attained and therefore, taking no corrective action; and

sending a subsequent request to said one server, determining that said one server did not successfully handle said request within said specified time period, determining that said number has been attained and therefore, taking corrective action.

11. A method as set forth in claim 10 wherein said corrective action is to remove said one server from said cluster or not send additional requests to said one server.

12. A method as set forth in claim 10 wherein said corrective action is to attempt to restart said one server.

13. A method as set forth in claim 10 wherein said corrective action is to automatically send a command to restart said one server.

14. A method for monitoring a plurality of servers in a cluster and taking corrective action for said servers, said method comprising the steps of:

setting a threshold equal to an integer greater than one;

sending a request to one of said servers, determining that said one server did not successfully handle said request within a predetermined amount of time, incrementing a count, comparing said count to said threshold, determining that said count is less than said threshold and therefore, taking no corrective action; and

sending another request to said one server, determining that said one server did not successfully handle said request within said predetermined amount of time, incrementing said count, comparing said count to said threshold, determining that said count equals or exceeds said threshold and therefore, taking corrective action.

15. A method as set forth in claim 14 wherein said corrective action is to remove said one server from said cluster or not send additional requests to said one server.

16. A method as set forth in claim 14 wherein said corrective action is to automatically send a command to restart said one server.

17. A method as set forth in claim 14 wherein said corrective action is to automatically request a memory dump from said one server and notifying a systems administrator or operator of said memory dump.

18. A method for monitoring a plurality of servers in a cluster and taking corrective action for said servers, said method comprising the steps of:

sending a request to one of said servers, and determining if said one server successfully handles said request within a predetermined time period; and

if said one server does not successfully handle said request within said predetermined time period or does not respond that it handled said request within said predetermined time period, notifying a dispatcher for said one server to remove said one server from said cluster or not to send any subsequent requests to said one server, and automatically sending a request to said one server to restart said one server.

19. A method as set forth in claim 14 wherein said other request is for different information than the first said request.